Table of content:

– Abstract

Problem statement: - Given an English sentence, translate it into Spanish.

Usage: - Translations have always been an effective tool to convey the meaning, tone, and intention of a message displayed in a particular language. Politically speaking it also narrows down the barriers between diverse cultures and regions. However, in this technology driven world, where scribbling pages is no longer considered effective over scrolling screens, NMT helps automate these translations.

Cheaper than human translation system and more flexibility over multiple languages.

• Literature survey

The Research papers......

1. Neural machine translation: A review of methods, resources, and tools (Tan et al, 2020)
2. Sequence to Sequence Learning with Neural Networks (Sutskever, Vinyals, V. Le, 2014)
3. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches (Cho, Merrienboer, Bahdanau, and Bengio, 2014)
4. LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection (Malhotra et al, 2016)
5. Neural Machine Translation with Supervised Attention (Liu, Utiyama, Finch and Sumita, 2016)

- ## Approach

About Algorithms and Working Methods

- **Dataset source**

  The data was open source available at https://www.manythings.org/anki/spa-eng.zip.

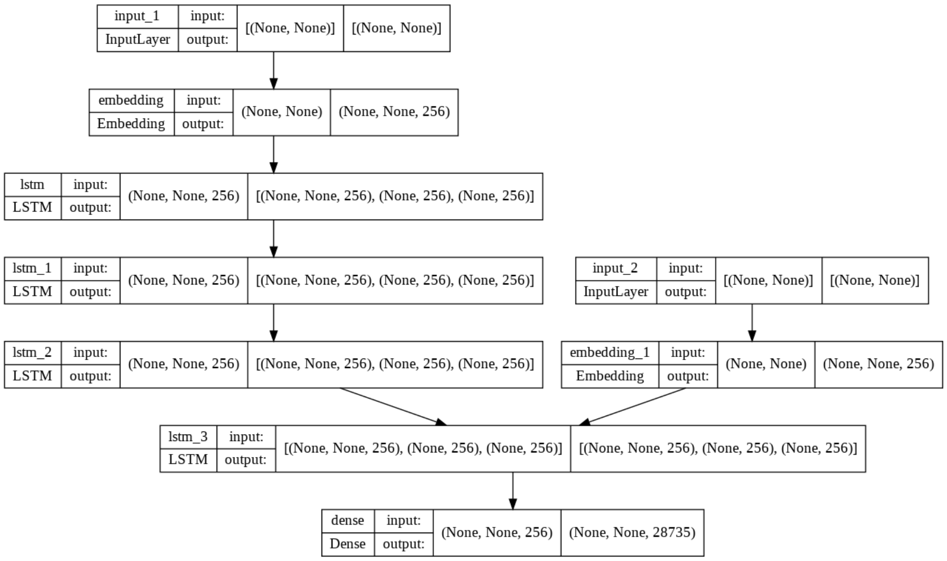
- **Data description**

  The data was present in .txt format which contained of English sentences and their corresponding Spanish text.

- **Data cleaning**
  1. It was converted into a data frame
  2. The texts were lower-cased
  3. Punctuations were removed
  4. Extra white spaces were stripped off

- **NLP Concepts**
  1. A vocabulary of words was created and sorted alphabetically of both source and target languages.
  2. The sentences were padded based on the maximum length.

- **Model Architecture**
  a. Encoder input -> embedding -> LSTM1 -> LSTM2 -> LSTM3 -> decoder input -> decoder embedding -> decoder LSTM -> decoder dense (SoftMax)
  b. Follows the principle of teacher forcing while training the model.


input_1 | InputLayer | input: | output: | [(None, None)] | [(None, None)]
embedding | Embedding | input: | output: | (None, None) | (None, None, 256)
lstm | LSTM | input: | output: | (None, None, 256) | [(None, None, 256), (None, 256), (None, 256)]
lstm_1 | LSTM | input: | output: | (None, None, 256) | [(None, None, 256), (None, 256), (None, 256)]
input_2 | InputLayer | input: | output: | [(None, None)] | [(None, None)]
lstm_2 | LSTM | input: | output: | (None, None, 256) | [(None, None, 256), (None, 256), (None, 256)]
embedding_1 | Embedding | input: | output: | (None, None) | (None, None, 256)
lstm_3 | LSTM | input: | output: | [(None, None, 256), (None, 256), (None, 256)] | [(None, None, 256), (None, 256), (None, 256)]
dense | Dense | input: | output: | (None, None, 256) | (None, None, 28735)

- **State of the art**

  - As surveyed in the research papers, we would need a method that is domain independent and learns to map sequences into sequences to handle sequential nature of sentences.
  - A many-to-many RNN (Recurrent Neural Net) would be used for the purpose which would consist of embedding, encoder, decoder, and classification layer.

- The input sentence would be word embedded, encoded into vector whose combined output would be fed as input to the decoder layer which extracts necessary information from the encoder layer.

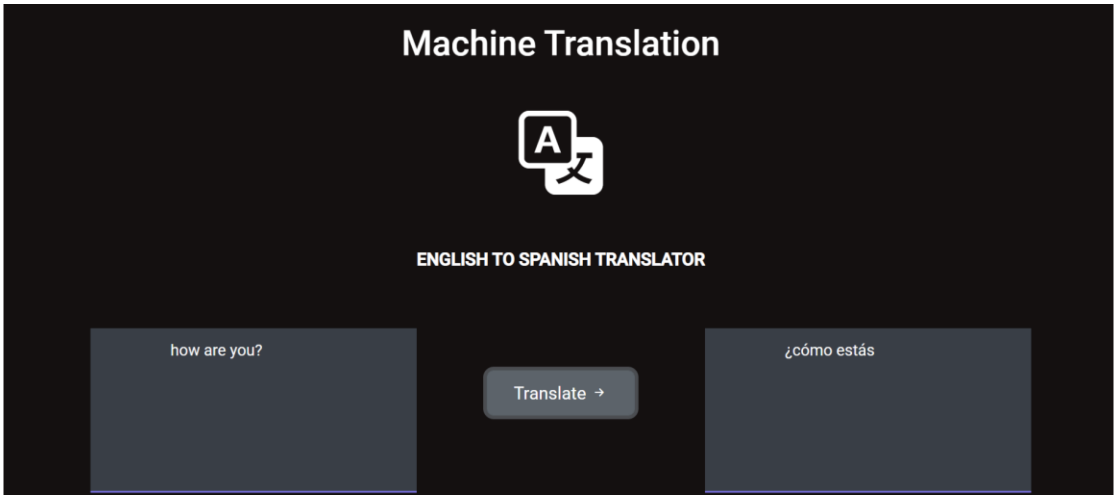
- Then the decoder is converted into a SoftMax dense layer, which ensures that our output is a valid probability.
- In long sentences, there are long term dependencies. To tackle this attention mechanism was introduced that works based on relevance between certain keys and values.
- RNNs (Recurrent neural networks) are one of the most powerful and complex neural networks, consequently facing vanishing gradient problem, which is handled by LSTMs (Long Short-Term Memory), which functions based on update, forget, and output gate.

- **The Adam optimizer was used, and the model was trained with a batch size of 128 for 12 epochs to give an accuracy of around 85%.**

- **The ML (Machine Learning) model was then scripted into an interactive website using Flask framework and is yet to be deployed on a cloud platform.**

- **Results and examples**

–



Machine Translation

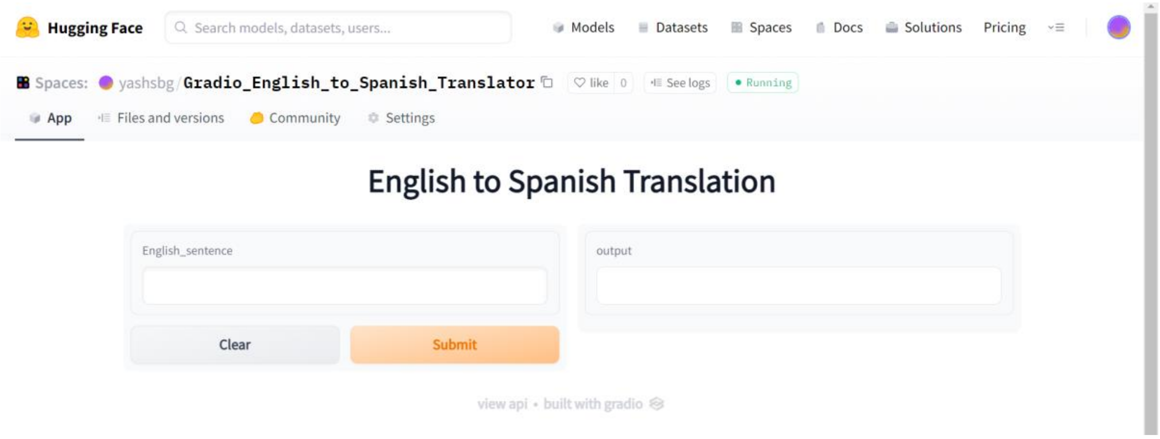
ENGLISH TO SPANISH TRANSLATOR

how are you?

¿cómo estás

- **There is also an open-source pre-trained library for NMT, called HuggingFace, which was deployed into a permanent website using gradio web application and HuggingFace spaces.**

–



English to Spanish Translation

English_sentence

output